Multi-Modal Word Synset Induction

Jesse Thomason and **Raymond J. Mooney**
Department of Computer Science, University of Texas at Austin
Austin, TX 78712, USA
{jesse, mooney}@cs.utexas.edu

Abstract

A word in natural language can be polysemous, having multiple meanings, as well as synonymous, meaning the same thing as other words. Word sense induction attempts to find the senses of polysemous words. Synonymy detection attempts to find when two words are interchangeable. We combine these tasks, first inducing word senses and then detecting similar senses to form word-sense synonym sets (*synsets*) in an unsupervised fashion. Given pairs of images and text with noun phrase labels, we perform synset induction to produce collections of underlying concepts described by one or more noun phrases. We find that considering multi-modal features from both visual and textual context yields better induced synsets than using either context alone. Human evaluations show that our unsupervised, multi-modally induced synsets are comparable in quality to annotation-assisted ImageNet synsets, achieving about 84% of ImageNet synsets' approval.

1 Introduction

Semantic understanding in language is complicated by polysemous words that have multiple, distinct meanings, and by synonymous sets of words that have the same underlying meaning. The word "bank," for example, has at least two distinct meanings: a financial institution and the edge of a river. Manually constructed lexical resources such as WordNet [Fellbaum, 1998] organize noun phrase meanings into senses which can be taken on by one or more noun phrases. Sets of synonymous senses are called *synsets*. For example, one WordNet synset contains both "bank" and "depository financial institution," two noun phrases that refer to the same underlying meaning.

The ImageNet [Deng *et al.*, 2009] corpus provides images that can be used as visual context for a subset of WordNet synsets. ImageNet required extensive annotation to construct, is limited to its current coverage, and is only available in English. In this work, we introduce *multi-modal word synset induction*, which automatically creates an ImageNet-like resource from a raw collection of images and associated texts annotated with noun phrases. The only initial annotation required is an association between noun phrases and observations, and our method produces synsets without further supervision.

Word sense induction (WSI) automatically determines the senses of a word [Pedersen and Bruce, 1997]. Text-based WSI is well-studied and discovers senses by clustering a word's textual contexts. The multiple senses for "bank" can be recognized as two clusters: one near words like "money" and "deposit"; and another near words like "river" and "shore." Word similarity tasks attempt to discover words with related meanings. Performing this kind of similarity search over word senses, we can discover synsets. To our knowledge, this work is the first to chain polysemy detection via WSI and synonymy detection through sense similarity to induce synsets.

Other notions of context, such as images a word is used to describe, can also be used to discover word senses. For instance, the two readings of "bank" are both textually and visually distinct. When detecting polysemy via WSI and synonymy through similarity, we consider both textual and visual contexts for noun phrases.

For this new task, we construct and release a corpus of images paired with web text, each labeled with a noun phrase, from ImageNet synsets, and induce synsets automatically from these. We use the WSI metrics from the SemEval-2010 Word Sense Induction and Disambiguation task [Manandhar *et al.*, 2010], which evaluate systems performing WSI, to measure the quality of the induced synsets against the gold standard from ImageNet. Additionally, we gather human judgments about the quality of induced synsets and ImageNet synsets.

A multi-modal approach using visual and textual features outperforms uni-modal approaches to synset induction in both automated and human evaluations. Human judgments rate our synsets from multi-modal induction as sensible about 84% as often as ImageNet's, suggesting that our unsupervised synsets are comparable in understandability to human-constructed ones.

2 Related Work

In distributional semantics, learning a single vector for an ambiguous word results in a representation that averages that word's ambiguous senses. First identifying senses and then

Noun phrase relationships			
synonymous	polysemous	both	neither
4019	804	1017	2586

Table 1: Number of noun phrases that are synonymous, polysemous, both, or neither in the ImageNet synsets S used in our experiments.

producing separate vectors for each sense has been shown to improve the performance of models of distributional semantics [Reisinger and Mooney, 2010]. Word sense induction is typically approached from distributional, textual context [Pedersen and Bruce, 1997; Schutze, 1998; Bordag, 2006; Manandhar *et al.*, 2010; Di Marco and Navigli, 2013]. We go beyond sense induction to additionally group similar senses into synsets, and we use both visual and textual observations of noun phrases to do so.

Other work has used visual information to disambiguate word senses, but assumes the senses of each word are known in advance [Barnard and Johnson, 2005]. Using both textual and visual information to perform WSI has been done on datasets where every input word is known in advance to be polysemous [Loeff *et al.*, 2006; Saenko and Darrell, 2008]. By contrast, our data contains polysemous, synonymous, and monosemous noun phrases. Additionally, we perform an explicit synonymy detection step to create synsets out of induced word senses, unlike other multi-modal word sense work [Lucchi and Weston, 2012]. Our synonymy detection step is related to lexical substitution [McCarthy and Navigli, 2007], but at the word sense level.

Similar works use co-clustering in separate textual and visual spaces, treating textual clusters as word senses and visual clusters as lower-level iconographic senses (such as different viewpoints for or orientations of an object) [Chen *et al.*, 2015]. We use deep image features from the VGG network [Simonyan and Zisserman, 2014] trained for object recognition, which removes the need for iconographic distinctions. Some work uses images and text to discriminate between word senses, but takes multiple senses as known, rather than inducing them automatically [Kanishcheva and Angelova, 2016].

The VGG network is trained from a subset of synsets and their associated images in ImageNet to take an image as input and identify the synset it belongs to. We hold out the synsets used to train VGG as validation data in our work. Other recent work has used the VGG network to extract visual features from objects [Thomason *et al.*, 2016], for developing similarity metrics within ImageNet [Deselaers and Ferrari, 2011], and for lexical entailment detection [Kiela *et al.*, 2015].

3 Dataset

We selected a subset of ImageNet synsets that were leaves in the WordNet hierarchy (e.g. “kiwi” but not “animal”) and were not used to train the VGG network. Table 1 gives the number of noun phrases which participated in polysemous and synonymous relationships among these 6,710 synsets, S.

We took the synsets used to train the VGG network as a development data set, V. We performed reverse image search[1] to get text from web pages on which images in V's synsets appeared. Images for which too little text could be extracted were removed from consideration. We performed latent semantic analysis (*LSA*) [Deerwester *et al.*, 1990] on term frequency-inverse document frequency (*tf-idf*) vectors of bag-of-words representations of this text to create a 256-dimensional text feature space.[2]

For each synset $s \in S$, deep visual features and LSA embedding features were extracted for up to 100 images per noun phrase associated with s in ImageNet. We arbitrarily selected the first $100 * |s|$ valid image URLs listed for the synset by the ImageNet API, eliminating ones for which too little text data could be found via reverse image search. Visual features were the activations of the 4,096-dimensional, penultimate layer of the VGG network given the image as input. This yielded a set of image observations I_s.

For each image, we gathered web text (about 400 words per image) as above and embedded it in our LSA space to get textual observations T_s. We expect text observations to be sense-specific for the images they are paired with, since, for example, a web page with a picture of a bank building is unlikely to discuss rainfall and flash floods. So, for each s, up to 100 multi-modal observations $O_s = \langle I_s, T_s \rangle$ are available per noun phrase. We make this corpus of ImageNet synsets associated with text, VGG features, and LSA embedding features per synset observation URL available.[3]

For each noun phrase $np \in NP$, we associate observations with np from each synset in which it participated by dividing each O_s evenly among participating noun phrases (illustrated in Figure 1). We refer to noun phrase observations as O_{np}. We note that these even divisions may not reflect a realistic distribution of senses (i.e. the fruit sense of ‘kiwi’ dominates the bird and people senses), but different hyperparameters could be set for specific domain distributions.

4 Synset Induction Method

Given this corpus of noun phrase image and text observations, we perform polysemy-detecting WSI to induce senses followed by synonymy detection to form synsets (Figure 2).

We performed synset induction using only visual features, only textual features, and both. Our induction algorithms are based on clustering. Using an early fusion paradigm [Bruni *et al.*, 2014] with cosine distance to combine modalities, we calculate distance $d(o_1, o_2)$ between observations as follows:

$$\begin{aligned} d(o_1, o_2) =& (\alpha) cosd(I_{o_1}, I_{o_2}) + \\ & (1-\alpha) cosd(T_{o_1}, T_{o_2}), \\ cosd(a, b) =& 1 - \frac{a \cdot b}{\|a\| \|b\|}, \end{aligned}$$

where α controls the relative influence of visual (I) and textual (T) features. We perform vision-only, text-only, and

[1] https://github.com/vivithemage/mrisa

[2] In early experiments we tried Word2Vec [Mikolov *et al.*, 2013] embeddings trained over this development text associated with synsets of V, but achieved better performance from simple LSA, possibly due to the small size of the development corpus.

[3] https://github.com/thomason-jesse/synpol

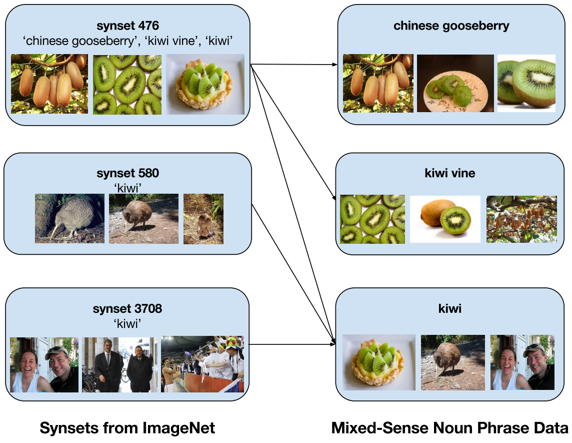


Figure 1: Noun phrase observations O_{np} (right) are extracted from ImageNet synsets (left). Our task is to automatically induce synsets from these noun phrase's image and text observations in an unsupervised fashion.

multi-modal induction by setting α to 1, 0, and 0.5, respectively.

Polysemy detection. k-means clustering powers polysemy detection, where k is estimated for each set of observations O_{np} using the gap statistic [Tibshirani *et al.*, 2001]. Intuitively, the gap statistic selects the smallest number of clusters k that reduces within-dispersion compared to $k-1$ by more than chance. Additionally, we enforce a constraint that no induced sense has fewer than 20 observations (estimated as the mean senses per noun phrase minus one standard deviation in the development data). Consequently, the observations O_{np} for each noun phrase np are clustered into k senses, yielding sense observation sets O_{np,k_i} for $k_i \in 0 \ldots k$. Together, these observation sets form a set of induced senses G.

Synonymy detection. Using the gap statistic to estimate an optimal number of clusters k^* for synonymy detection is inappropriate because we know k^* is on the order of $|G|$, since the number of synsets is much closer to the total number of word senses than to 1. The gap statistic is best applied when looking for a minimum optimal k^*, and further sensible divisions of k^* well-separated clusters may exist within these larger clusters [Tibshirani *et al.*, 2001].

Instead, we use a greedy merging approach. We compute a mean observation vector for each induced sense $O_{np,k_i} \in G$, as well as the pairwise distance $d(m_1, m_2)$ between all mean sense vectors. Greedy merges of the nearest means produces a final set of K induced synsets, R, each of which comprises no more than L distinct word senses.

Membership in each induced synset $r \in R$ is the union of observations of the senses $g_a \ldots g_b \in G$ whose observations were merged (i.e. $r = \cup\{g_a \ldots g_b\}$). K is set based on the ratio of senses to synsets in the development data V (so K fluctuates depending on the number of senses to be clustered). The maximum number of senses per synset, $L = 32$, is also estimated from V.

5 Experimental Evaluation

Both automated and human evaluations demonstrate that multi-modal synset induction outperforms uni-modal induction. More importantly, human judges do *not* significantly favor ImageNet synsets over multi-modal, induced synsets; however, humans *do* favor ImageNet's over uni-modally induced synsets.

Automated Evaluation. We computed the *v-measure* [Rosenberg and Hirschberg, 2007] of the induced synsets, calculated as the harmonic mean of their *homogeneity* and *completeness* with respect to the gold-standard ImageNet synsets. High homogeneity means the induced synsets mostly contain observations that correspond to a single gold synset, while high completeness means each gold synset's observations are mostly assigned to the same induced synset. We do not compare our word sense induction method to past WSI datasets [Manandhar *et al.*, 2010; Navigli and Vannella, 2013], because we take an additional synonymy detection step, and we consider textual and visual information jointly, while existing corpora use only text.

Homogeneity and completeness are defined in terms of the class entropies $H(S)$ and $H(R)$ of the gold-standard ImageNet synsets S, induced synsets R, and their conditional entropies $H(S|R)$ and $H(R|S)$. Specifically, homogeneity $h(S, R)$ is calculated as follows:

$$H(S) = -\sum_{i=1}^{|S|} \frac{\sum_{j=1}^{|R|} a_{ij}}{N} \log \frac{\sum_{j=1}^{|R|} a_{ij}}{N},$$

$$H(S|R) = -\sum_{j=1}^{|R|} \sum_{i=1}^{|S|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|S|} a_{kj}},$$

$$h(S,R) = \begin{cases} 1 & H(S) = 0 \\ 1 - \frac{H(S|R)}{H(S)} & H(S) > 0 \end{cases},$$

with a_{ij} the number of observations of gold synset S_i that ended up in induced synset R_j, and N the total number of observations in the dataset. Completeness $c(S, R)$ is defined as follows:

$$H(R) = -\sum_{j=1}^{|R|} \frac{\sum_{i=1}^{|S|} a_{ij}}{N} \log \frac{\sum_{i=1}^{|S|} a_{ij}}{N},$$

$$H(R|S) = -\sum_{i=1}^{|S|} \sum_{j=1}^{|R|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|R|} a_{ik}},$$

$$c(S,R) = \begin{cases} 1 & H(R) = 0 \\ 1 - \frac{H(R|S)}{H(R)} & H(R) > 0 \end{cases},$$

with the v-measure defined as the harmonic mean of $h(S, R)$ and $c(S, R)$.

We also computed the *paired f-measure* [Manandhar *et al.*, 2010], the harmonic mean of the paired precision and recall between the ImageNet and induced synsets. Rather than count membership overlap between two sets, paired f-measure compares membership overlap between sets of sets.

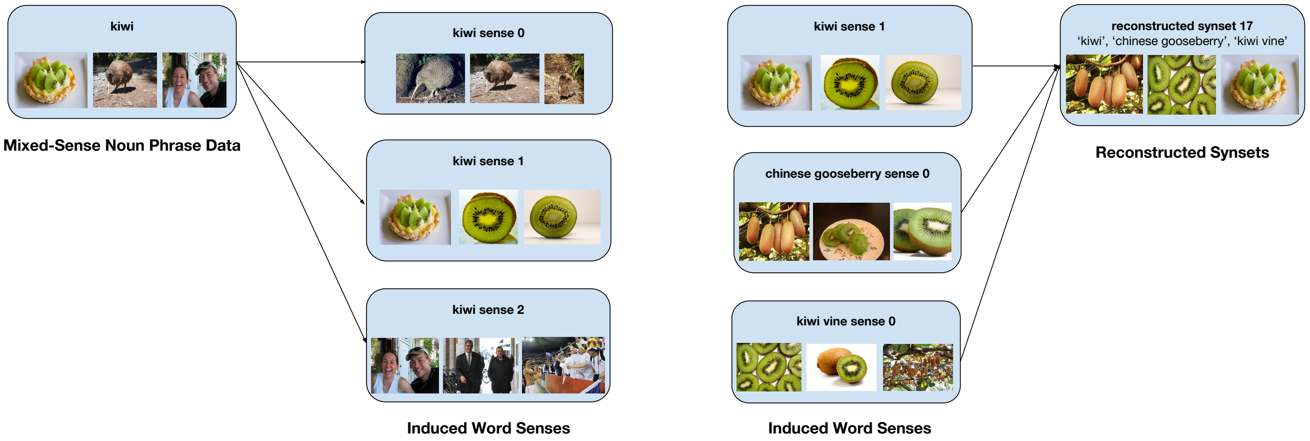


Figure 2: **Left:** We induce senses for each noun phrase by clustering among its observation feature vectors. **Right:** We induce synsets by calculating the mean sense observation vectors across induced senses, then clustering over those means.

Specifically, we count the number of observation pairs (o_i, o_j) that are members of both synset s and induced synset r to get an overlap score between each $s \in S$ and $r \in R$. There are $\binom{|s|}{2}$ observation pairs for each s and $\binom{|r|}{2}$ observation pairs for each r, across all such s and r comprising $C(S)$ gold pairs and $C(R)$ induced pairs, respectively. Then paired f-measure $f(S, R)$ is defined as the harmonic mean of paired precision $p(S, R)$ and recall $r(S, R)$:

$$p(S, R) = \frac{|C(S) \cap C(R)|}{|C(R)|},$$
$$r(S, R) = \frac{|C(S) \cap C(R)|}{|C(S)|}$$

Results are presented in Table 2. The multi-modal approach achieves the highest v-measure and is tied for highest paired f-measure. This unsupervised task operates over a whole dataset, not a train/test split where cross validation could be performed, so we have simply highlighted the highest score for each metric. These paired scores are low compared to those of strict word sense induction [Manandhar *et al.*, 2010] because our method attempts to induce synsets, not just word senses, adding another layer of difficulty to the problem.

Homogeneity and paired precision are maximized when every observation has its own synset. Completeness and paired recall are maximized when all observations belong to a single synset. The vision-only system overproduces synsets and increases precision, while the text-only system underproduces synsets and increases recall. The multi-modal system is able to balance between these flaws to achieve high v-measure and paired f-measure.

Human Evaluation. We observed that ImageNet synsets do not necessarily match human-distinguishable categories. For example, ImageNet distinguish photos of Croatian from Ukranian peoples, as well as having a synset for people who could be described as "energizers." By contrast, multi-modal induction grouped senses of noun phrases referring to people together in one large synset.

We created an Amazon Mechanical Turk task to evaluate the quality of our induced synsets according to human judgment. Given a noun phrase and a set of synsets that noun phrase participated in, annotators marked whether they thought the sets were 'more sensible' or 'more confusing.' Figure 3 shows the interface with one of the validation examples, discussed below.

Annotators were walked through three examples of how the word "bank" might be split into synsets before the task began. Two senses containing bank (financial institution and riverbank) were shown, with one 'sensible' example of them well-separated in two synsets, one 'confusing' example of the senses lumped together in a single synset, and one 'confusing' example where the senses were separated but there were two distinct synsets for financial institutions even though this is a single concept.

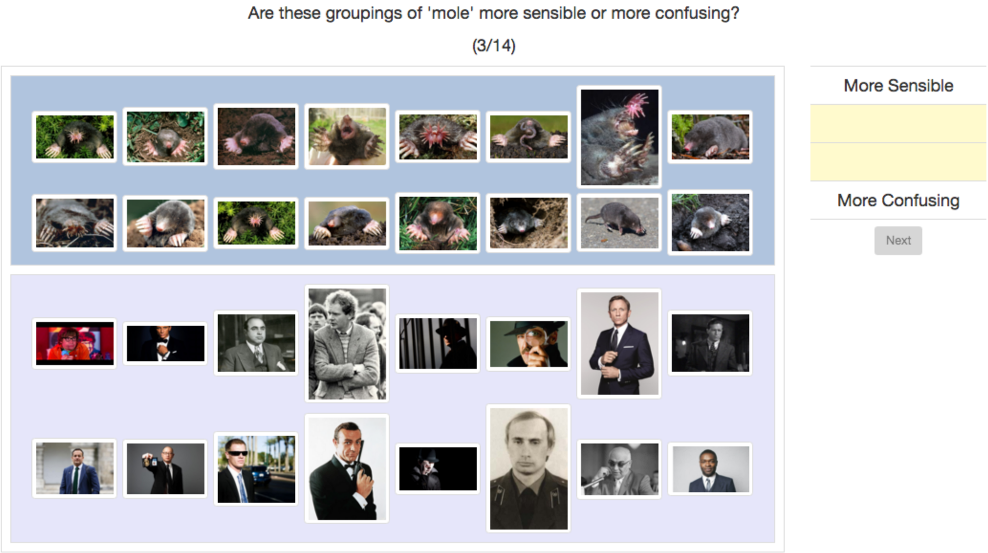
Three noun phrases were selected randomly from the corpus for each annotator. Annotators evaluated vision-only, text-only, and multi-modal induced synsets, as well as the gold standard ImageNet synsets, that the noun phrases participated in. The ordering of the 12 sets (3 noun phrases × 4 models) was randomized before being shown to the annotator. Two hand-created validation examples[4] were inserted in random positions, and data from users who answered either of these incorrectly was discarded.

After removing data from users who failed validations (nearly half did—the task is challenging for Mechanical Turkers), and noun phrases assigned to multiple annotators who did not reach consensus (e.g. tie in whether sets of synsets are sensible or confusing), we had 156 noun phrases annotated across all four models (624 annotator decisions total) by 58 distinct annotators. We calculated the average annotator decision per noun phrase/model combination ('more sensible'$= 1$, 'more confusing'$= 0$), and averaged those decisions across noun phrases to get human scores per model,

[4] One of well-separated "mole" senses (animal and spy), and one of incorrectly grouped "crane" senses (birds and construction)

	|synsets|	**h**	**c**	**v**	**p**	**r**	**f**	**human**
ImageNet	6710	1.0	1.0	1.0	1.0	1.0	1.0	0.470
vision	9976	**0.897**	0.888	0.893	**0.326**	0.440	**0.375**	0.388
text	6406	0.853	**0.911**	0.881	0.173	0.496	0.256	0.346
vision+text	8216	0.887	0.910	**0.899**	0.286	**0.543**	**0.375**	**0.395**

Table 2: Homogeneity (**h**), completeness (**c**), v-measure (**v**), paired precision (**p**), recall (**r**), f-measure (**f**), and **human** scores of our induced synsets with visual features only, textual features only, and both. Bold indicates highest value by modality (excluding the gold-standard ImageNet).


Are these groupings of 'mole' more sensible or more confusing?

(3/14)

Figure 3: The Mechanical Turk interface used to gather annotations. The noun phrase "mole" was a hand-crafted validation example of 'more sensible' synsets—one for the burrowing animal and one for the spy.

shown in Table 2.

ImageNet synsets are only rated more sensible than confusing about half the time in our sample, highlighting the noisiness of ImageNet synsets. We conducted paired (by noun phrase) Student's t-tests between all models and found that only the differences between ImageNet and the uni-modal models are significant ($p < 0.05$). Humans found multi-modal induced synsets sensible about 84% as often as ImageNet synsets ($.470 \cdot .84 \approx .395$), without requiring explicit annotations to build synsets from noun phrases and observations.

Figure 4 shows an example where annotators favored our multi-modal, induced synsets versus ImageNet. The patterns of vision-only induction overproducing synsets (e.g. two senses of "washboard, splashboard" for the presence and absence of a human) while text-only induction under-produces them (e.g. combines "washboard" and "dulcimer" instruments in one synset) are common. Multi-modal induction's advantage lies in balancing these opposing trends, producing more coherent synsets like the two shown for "washboard."

For other noun phrases, like "amphitheater," ImageNet distinguishes it from "coliseum" while unsupervised induction recognizes their similarity, and human annotators agree with collapsing the two. Situations like this one, where ImageNet makes a distinction human annotators disagree with, is also common among groups of peoples. For example, ImageNet separates nationalities like "Austrian" and "Croation", while automatic induction (across modalities) favors putting groups of people together without respect to nationality.

6 Conclusions and Future Work

We introduce the task of *multi-modal word synset induction* and an automatic method to construct synsets from image and text observations labeled with noun phrases. Additionally, we create a dataset of image and text feature observations, drawn from ImageNet and reverse image search and processed by the VGG network and LSA, labeled with noun phrases from ImageNet.

We show that a multi-modal, unsupervised clustering approach in which visual and textual features are considered together outperforms uni-modal clustering at the synset induction task both quantitatively and qualitatively. Human annotators rate our multi-modal, induced synsets sensible 84% as often as gold-standard ImageNet's, suggesting our unsupervised method is competitive with manual annotation for creating synsets from noun phrase-level observations.

By applying these methods to a set of images labeled with (possibly polysemous and synonymous) nouns in a language

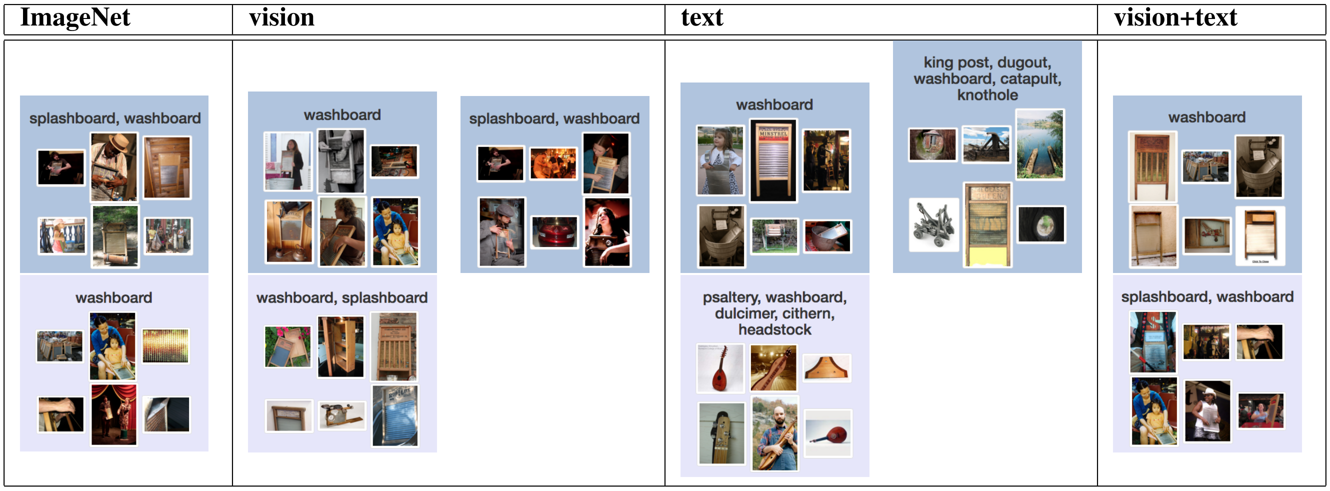


Figure 4: Human annotators favored the multi-modal, induced synsets for noun phrase "washboard" over ImageNet's and other models' synsets. ImageNet fails to properly distinguish the "washboard" senses of a household object and instrument, vision alone creates too many instrument senses, and text alone overgeneralizes the instrument sense. Multi-modal induction properly separates the household object and instrument senses.

other than English,[5] or in a specific domain, new ImageNet-like resources of synsets could be induced rather than crafted by hand.

Other future directions could examine alternative clustering methods for both polysemy and synonymy detection. For example, designing a non-parametric form of hierarchical agglomerative clustering may be appropriate for synonymy detection. Additionally, varying the weighting parameter α between modalities rather than using an equal weight may reveal a more effective balance.

Our methods can be applied to any vector representation of instances labeled with discrete classes that need to be disambiguated. For example, in a grounded language system where word meanings are associated with real-world object properties in visual [Perera and Allen, 2013; Parde *et al.*, 2015] or multi-modal space [Thomason *et al.*, 2016], instances are object representations and labels are adjectives and nouns applied to those objects. Words like "round" are visually polysemous, since something can be flat and circular or spherical and still be called "round." This work could tease out these meanings of "round" and subsequently join the meaning of "spherical" to the appropriate one. Additionally, discovering that "light" is polysemous across modalities (coloration versus weight) and joining the color sense to "bright" and the weight sense to "lightweight" could make robot-human communication clearer, since an embodied agent should prefer the less polysemous descriptor words when describing things to a human.

Multi-modal representations could be used for less categorical labels than noun phrases, such as abstract action concepts like "grasp" or "lift". In those cases, feature representations of actions might be videos of actions, or, in the case of robotic manipulation, haptic and proprioceptive feedback. Synset induction on these action concepts may, for example, reveal different senses of "grasp" for finger positions as well as joining senses of "grasp" and "clench" into a synset for a tight grip.

Acknowledgments

We would like to thank our anonymous reviewers for their feedback and insights and Subhashini Venugopalan for her help in engineering deep visual feature extraction. This work is supported by a National Science Foundation Graduate Research Fellowship to the first author, an NSF EAGER grant (IIS-1548567), and an NSF NRI grant (IIS-1637736).

References

[Barnard and Johnson, 2005] Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. *Journal of Artificial Intelligence*, 167:13–30, 2005.

[Bordag, 2006] Stefan Bordag. Word sense induction: Triplet-based clusering and automatic evaluation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 137–144, 2006.

[Bruni *et al.*, 2014] Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.

[Chen *et al.*, 2015] Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom Mitchell. Sense Discovery via Co-Clustering on Images and Text. In *Computer Vision and Pattern Recognition*, 2015.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] For example, utilizing aligned WordNets, since different languages have different polysemous and synonymous word tokens `http://compling.hss.ntu.edu.sg/omw/`

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition*, 2009.

[Deselaers and Ferrari, 2011] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition*, pages 1777–178, June 2011.

[Di Marco and Navigli, 2013] Antonio Di Marco and Roberto Navigli. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754, 2013.

[Fellbaum, 1998] Christiane D. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Kanishcheva and Angelova, 2016] Olga Kanishcheva and Galia Angelova. *About Sense Disambiguation of Image Tags in Large Annotated Image Collections*, pages 133–149. Springer International Publishing, 2016.

[Kiela *et al.*, 2015] Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 119–124, Beijing, China, July 2015.

[Loeff *et al.*, 2006] Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. Discriminating image senses by clustering with multimodal features. In *Proceedings of the Conference on Computational Linguistics*, pages 547–554, Stroudsburg, PA, USA, 2006.

[Lucchi and Weston, 2012] Aurelien Lucchi and Jason Weston. Joint image and word sense discrimination for image retrieval. In *European Conference on Computer Vision*, pages 130–143. Springer, 2012.

[Manandhar *et al.*, 2010] Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 63–68, Stroudsburg, PA, USA, 2010.

[McCarthy and Navigli, 2007] Diana McCarthy and Roberto Navigli. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics, 2007.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada, 2013.

[Navigli and Vannella, 2013] Roberto Navigli and Daniele Vannella. Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *In Second Joint Conference on Lexical and Computational Semantics*, 2013.

[Parde *et al.*, 2015] Natalie Parde, Adam Hair, Michalis Papakostas, Konstantinos Tsiakas, Maria Dagioglou, Vangelis Karkaletsis, and Rodney D. Nielsen. Grounding the meaning of words through vision and interactive gameplay. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1895–1901, Buenos Aires, Argentina, 2015.

[Pedersen and Bruce, 1997] Ted Pedersen and Rebecca Bruce. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language*, pages 197–207, 1997.

[Perera and Allen, 2013] Ian Perera and James F. Allen. Sall-e: Situated agent for language learning. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1241–1247, Bellevue, Washington, USA, 2013.

[Reisinger and Mooney, 2010] Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, 2010.

[Rosenberg and Hirschberg, 2007] Andrew Rosenberg and Julia Hirschberg. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, 2007.

[Saenko and Darrell, 2008] Kate Saenko and Trevor Darrell. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems 21*, pages 1393–1400. 2008.

[Schutze, 1998] Hinrich Schutze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository*, abs/1409.1556, 2014.

[Thomason *et al.*, 2016] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond Mooney. Learning multi-modal grounded linguistic semantics by playing "I spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3477–3483, July 2016.

[Tibshirani *et al.*, 2001] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.